Method and System for Comparing Information Contents

BACKGROUND OF THE INVENTION

FIELD OF THE INVENTION

The invention relates to information processing. More particularly, the invention relates to a system and a family of methods that provide for fast and reliable comparison of information contents.

DESCRIPTION OF RELATED TECHNOLOGY

An organization may receive thousands of emails every day. The received emails may be automatically stored in a relational database from which customer service representatives may retrieve, read, and act upon. For various reasons, some malicious, some by mistake, others due to errors in the infrastructure, a number of duplicate copies of an email may be received or stored in the relational database.

There are many problems with storing duplicate copies of an email. Storing large number, sometimes thousands, of identical email in a database severely affects the system performance, and wastes personnel time. Since the received emails are typically large in size, they are usually stored as Binary Large Objects (BLOBs). The BLOBs are not searchable for determining whether they include any duplicates, and even if they were searchable, it would be prohibitively time consuming. That is because the emails have to be stored in the relational database before being searched, and the existing search techniques are limited to the size and type of data to be searched.

There is a need, therefore, for detecting duplicate emails, before storing them in the system, in a fast and reliable way.

SUMMARY OF THE INVENTION

One presently preferred embodiment of the invention provides a system and a method for detecting whether received information content is identical to a plurality of stored information contents. The system and method may include the steps of determining a plurality of parameters, each representing one of the plurality of stored information contents, and storing the plurality of parameters. After receiving an information content, the system and method may determine a parameter representing the received information content, compare the parameter representing the received information content with the plurality of stored parameters; and indicate that the received information content is identical to a stored information content if the corresponding parameters are equal. In one embodiment, the parameters may be determined based on an order and a value of each character in the corresponding information content.

Another presently preferred embodiment of the invention provides a system and a method for comparing a plurality of information contents. The system and method may include the steps of determining a plurality of parameters, each representing one of the plurality of information contents, and comparing the plurality of parameters, such that equality between a pair of the plurality of parameters indicates that a corresponding pair of the plurality of information contents are identical.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a representation of an exemplary process for comparing information contents according to one embodiment of the invention; and

Figure 2 shows a representation of an exemplary system for implementing the process described in Figure 1.

DETAILED DESCRIPTION OF THE INVENTION

One embodiment of a process for comparing information contents is represented in Figure 1. The process may be applied to comparison of contents of any type of electronic files, databases, or data objects and constructs, including emails, Web pages, and the like. In the following, however, an exemplary discussion of the process for comparing information contents according to one embodiment of the present invention is presented in reference to emails. When an email is received by an organization's computing network, before storing the email, it is desirable to compare the content of the received email to the content of the previously received and stored emails to determine whether a duplicate copy of the received email is already stored in the system.

Since the received emails are typically large in size, they are usually stored as Binary Large Objects (BLOBs). The BLOBs are not searchable for determining whether they include any duplicates, and even if they were searchable, it would be prohibitively time consuming. That is because the emails have to be stored in the relational database before searching them, and the existing search techniques are limited to the size and type of data to be searched. According to one embodiment of the invention, a parameter that uniquely represents the content of each email may be determined for each received email, and the comparison process may be efficiently carried out on the parameters, rather than on the actual email contents. This process makes the comparison fast and reliable, and improves system performance and the personnel effectiveness.

Referring to Figure 1, after information content, *e.g.* an email, is received 102, through the Internet or any global communications network, the process determines 104 a parameter that uniquely represents the content of the received email. In step 106, the parameter representing the received email is compared with the previously stored parameters representing the previously received and stored emails. In one embodiment, the parameters may be single numerical values, which may be efficiently compared together by simple comparison techniques, thereby avoiding comparison of large size email contents.

In one embodiment, the parameter representing the content of an email may be determined using the following formula:

$$R=\left| n\,(\Sigma n^{0.1}a^{0.1})-(\Sigma n^{0.1})(\Sigma a^{0.1})/SQRT\,\{[n\,(\Sigma(n^{0.1})^{2})-\Sigma(n^{0.1})^{2}]\;[n\,(\Sigma(a^{0.1})^{2})-\Sigma(a^{0.1})^{2}]\}\right|$$

In the above formula, “R” stands for the parameter that uniquely represents the content of an email. The numerical value of “R” may be within zero and one. The factor “n” represents the position order of the constituent characters of the email, and the factor “a” represents a unique value for the constituent characters in the email. In one embodiment “a” may be represented by an ASCII code, but other codes may be used.

Table 1 shows some exemplary short information contents along with the corresponding unique “R” values. The “R” values shown in Table 1 are determined using ASCII values for the constituent characters of each information content, with precision of eight digits. A typical email may include up to several thousands of characters, and the corresponding “R” value may be determined with higher precision for higher accuracy.

Table 1

Information Content	ASCII values	“R” value
aaa	97, 97, 97	0.99878402
aaaa	97, 97, 97, 97	0.99867733
aab	97, 97, 98	0.99879121
bbb	98, 98, 98	0.99877977
xxy	120, 120, 121	0.99869948

If the result of parameter comparison in step 106 indicates that the parameter representing the received email is equal to one of the previously stored parameters, indicating that the received email is identical to one of the previously stored emails, the received email is not stored. On the other hand, if the result of parameter comparison in step 106 indicates that the parameter representing the received email is not equal to any of the previously stored parameters, indicating that the received email is not identical to any one of the previously stored emails, the received email may be stored in step 108, and the corresponding parameter may be stored in step 110.

The invention contemplates a new and unique system and a family of methods for comparing contents of information objects, such as emails, which may be implemented in a network of computer systems, interconnected by a global communications network, such as the Internet. A computer system may include user terminals, storage devices, processing units, input and output devices, and networking devices and software modules.

Figure 2 shows a representation of an exemplary system for implementing the different embodiments of the invention. The user terminals 202, 204 may include the hardware and software modules to implement the disclosed invention. The user terminals may also include the necessary devices and software modules to connect to the global telecommunication network 206, which may include the Internet. The information contents and the corresponding parameters may be maintained in the databases 208, 210.

The method and system disclosed herein provide for detecting duplicate information contents such as emails, before storing them in the system, in a fast and reliable way. Although the invention has been described in detail with reference to particular preferred or exemplary embodiments, persons possessing ordinary skill in the art to which this invention pertains will appreciate that various modifications and enhancements may be made without departing from the spirit and scope of the claims that follow.